The ecologist's field guide to sequence-based identification of biodiversity

Creer, Simon; Deiner, Kristy; Frey, Serita; Porazinska, Dorota; Taberlet, Pierre; Thomas, W. Kelley; Potter, Caitlin; Bik, Holly M.

Published in:
Methods in Ecology and Evolution

DOI:
10.1111/2041-210X.12574

Publication date:
2016

Citation for published version (APA):
Creer, S., Deiner, K., Frey, S., Porazinska, D., Taberlet, P., Thomas, W. K., Potter, C., & Bik, H. M. (2016). The ecologist's field guide to sequence-based identification of biodiversity. *Methods in Ecology and Evolution*, *7*(9), 1008-1018. https://doi.org/10.1111/2041-210X.12574

Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2016, **7**, 1008–1018 doi: 10.1111/2041-210X.12574

REVIEW

The ecologist's field guide to sequence-based identification of biodiversity

Simon Creer[1,*], Kristy Deiner[2], Serita Frey[3], Dorota Porazinska[4], Pierre Taberlet[5], W. Kelley Thomas[6], Caitlin Potter[1] and Holly M. Bik[7]

[1]*Molecular Ecology and Fisheries Genetics Laboratory, School of Biological Sciences, Environment Centre Wales Building, Bangor University, Deiniol Road, Bangor, Gwynedd LL57 2UW, UK;* [2]*Eawag: Aquatic Ecology, Überlandstrasse 133, 8600 Dübendorf, Switzerland;* [3]*Natural Resources and the Environment, University of New Hampshire, Durham, NH 03824, USA;* [4]*Department of Ecology and Evolutionary Biology, University of Colorado at Boulder, Boulder, CO 80309, USA;* [5]*Laboratoire d'Ecologie Alpine, CNRS UMR 5553, Université Joseph Fourier, BP 43, F-38041 Grenoble Cedex 9, France;* [6]*Department of Molecular, Cellular, & Biomedical Sciences, University of New Hampshire, Durham, NH 03824, USA; and* [7]*Center for Genomics and Systems Biology, Department of Biology, New York University, New York, NY 10003, USA*

Summary

1. The past 100 years of ecological research has seen substantial progress in understanding the natural world and likely effects of change, whether natural or anthropogenic. Traditional ecological approaches underpin such advances, but would additionally benefit from recent developments in the sequence-based quantification of biodiversity from the fields of molecular ecology and genomics. By building on a long and rich history of molecular taxonomy and taking advantage of the new generation of DNA sequencing technologies, we are gaining previously impossible insights into alpha and beta diversity from all domains of life, irrespective of body size. While a number of complementary reviews are available in specialist journals, our aim here is to succinctly describe the different technologies available within the omics toolbox and showcase the opportunities available to contemporary ecologists to advance our understanding of biodiversity and its potential roles in ecosystems.

2. Starting in the field, we walk the reader through sampling and preservation of genomic material, including typical taxonomy marker genes used for species identification. Moving on to the laboratory, we cover nucleic acid extraction approaches and highlight the principal features of using marker gene assessment, metagenomics, metatranscriptomics, single-cell genomics and targeted genome sequencing as complementary approaches to assess the taxonomic and functional characteristics of biodiversity. We additionally provide clear guidance on the forms of DNA found in the environmental samples (e.g. environmental vs. ancient DNA) and highlight a selection of case studies, including the investigation of trophic relationships/food webs. Given the maturity of sequence-based identification of prokaryotes and microbial eukaryotes, more exposure is given to macrobial communities. We additionally illustrate current approaches to genomic data analysis and highlight the exciting prospects of the publicly available data underpinning published sequence-based studies.

3. Given that ecology 'has to count', we identify the impact that molecular genetic analyses have had on stakeholders and end-users and predict future developments for the fields of biomonitoring. Furthermore, we conclude by highlighting future opportunities in the field of systems ecology afforded by effective engagement between the fields of traditional and molecular ecology.

Key-words: biodiversity, DNA sequencing, metabarcoding, metagenomics, metatranscriptomics, molecular ecology

Community ecology and biodiversity assessment

A recent British Ecological Society supplement ('100 Influential Papers') (http://www.britishecologicalsociety.org/100papers/100InfluentialPapers.html#1) makes for inspirational reading, highlighting just some of the substantial contributions that the field of ecology has made to our understanding of the natural world. Notable papers focus on ecosystem biodiversity relationships, predicting change in communities according to ecological traits, understanding food web interactions, above-ground–below-ground relationships and assessing the effectiveness of management for the promotion of biodiversity. In combination with newer and rapidly developing fields such as macroecology and species distribution modelling, it is all too tempting to try and further define and test general processes in spatial community ecology and make predictions regarding environmental change. Community ecology is generally

*Correspondence author. E-mail: s.creer@bangor.ac.uk

affected by four broad processes: selection (fitness as a consequence of biotic/abiotic interactions), drift (stochastic changes), speciation (creation of new species) and dispersal (i.e. spatial movements); just as population genetics is affected by selection, drift, mutation and gene flow (Vellend 2010). Nevertheless, ecology is inherently more complicated than population genetics, since community ecology features the interactions of multiple evolving dependent variables (i.e. organisms) with each other and their environment in space and time (Vellend 2010). In order to make advances in community ecology, we must be able to quantify and understand the processes of selection, drift, speciation and dispersal by enhancing our understanding of alpha and beta diversity. If we can comprehensively characterize entire communities and their biotic/abiotic interactions, we will be in a position to develop the necessary modelling tools required to make systems ecology predictions associated with change (Segata *et al.* 2013). Nevertheless, many contemporary ecological studies do not take into account entire communities for obvious reasons. The challenges associated with the identification of taxonomically intractable communities, the volume and taxonomic breadth of diversity that often needs to be sampled (Creer *et al.* 2010) and the lack of resources (e.g. funding to support taxonomists) required to perform species identifications are immediate obstacles that spring to mind. In short, the job of community-wide assessment is large and difficult, and there are not enough skilled employees to complete the ongoing tasks.

Lessons from the microbial biosphere

Many of the problems associated with the quantification of unculturable microbial communities have been overcome by employing the new generation of DNA sequencing technologies (Loman *et al.* 2012) often referred to as high-throughput sequencing (HTS). Combined with coordinated local and global sampling campaigns, the standardized format of nucleic acid sequence data is now enabling us to gain previously impossible insights into the alpha and beta diversity of unseen or untraceable communities. Therefore, whether you are interested in the below-ground diversity of Central Park (Ramirez *et al.* 2014) or the formation of global Genomic Observatories (Davies, Field & The Genomic Observatories Network 2012), there are initiatives underway to join. Moreover, using HTS for the assessment of biodiversity has expanded from the microbial and micro-eukaryotic (Bik *et al.* 2012) biospheres to macro-communities (Ji *et al.* 2013). By focusing on a range of genetic source material (e.g. community-level or environmental DNA), habitats and spatial scales, we can now comprehensively characterize entire communities and begin to unpick their biotic/abiotic interactions. Referred to as 'transformative' (Baird & Hajibabaei 2012) technology, harnessing the information held in DNA potentially has the power to overcome many limitations of classical biodiversity assessment. A narrow taxonomic focus, potential subjectivity and the typically low throughput and labour-intensive nature of manual species

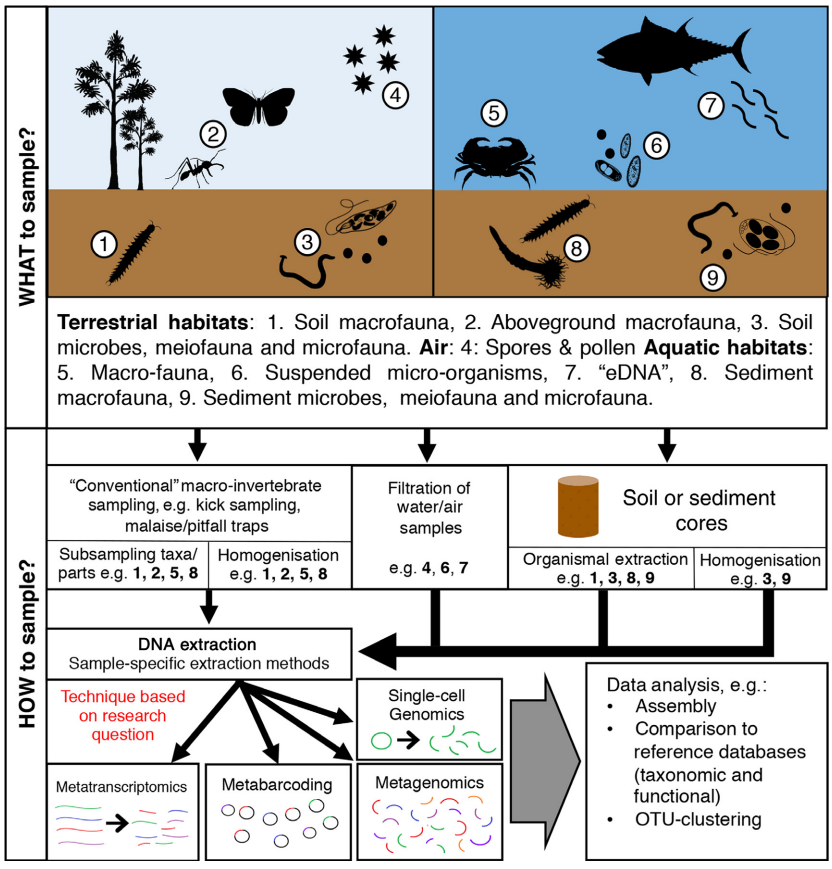

Fig. 1. Schematic showing the decisions involved in a molecular ecology workflow. Samples can be collected from a variety of different environments using appropriate collection techniques. DNA is then prepared and used to answer a variety of ecological questions: metabarcoding is used to answer questions about 'who' is present, while the function of communities or individuals can be established using a metagenomics, single-cell genomics or metatranscriptomics- see text for full details.

identification can generally all be overcome using HTS (Lawson Handley 2015). The purpose of this review therefore is to provide a succinct summary for ecologists of the different HTS approaches for the assessment of biodiversity (*sensu* genes to ecosystems; Fig. 1), identify case studies and showcase the ecological research opportunities afforded by contemporary DNA sequencing. A glossary of terms is provided in Box 1.

A brief history of molecular taxonomic identification

The use of taxonomically informative molecules has been key to establishing a phylogenetic framework for the vast uncharacterized biological diversity on earth. Early work focused primarily on genes encoding ribosomal subunits (rRNA) as universal 'orthologs', contributing to the phylogenetic understanding of prokaryotic life (Fox *et al.* 1980). Consequently, early studies of DNA isolated directly from environmental samples used these same molecules to place novel organisms into an evolutionary framework and to discover and confirm the extraordinary amount of biodiversity present in unculturable organisms from diverse environments (Giovannoni *et al.* 1990). While early work focused on bacteria and archaea, subsequent application of the same techniques to homologous molecules followed for microbial eukaryotes, where similar challenges of biodiversity discovery and underdeveloped taxonomies limited characterization of their diversity (Blaxter *et al.* 2005).

One of the early, transformative technical advances in environmental sequencing was the development of the polymerase chain reaction (PCR) (Saiki *et al.* 1988). The highly conserved sequences flanking phylogenetically informative regions found in rRNA loci paved the way for the rapid adoption of PCR-based amplification from environmental samples followed by cloning and sequencing of numerous clones. While most of the early studies on bacteria and archaea focused on the nuclear 16S rRNA gene, other taxonomic groups employed a diverse set of loci from the analogous eukaryotic rRNA gene array (e.g. ITS, 18S or 28S rRNA) (Bik *et al.* 2012; Epp *et al.* 2012), chloroplast genes (for plants) (Group *et al.* 2009) and mitochondrial DNA (for multicellular animals) in an attempt for species-specific resolution (Table 1).

While the advent of PCR made it possible to effectively sample organismal diversity directly from the environment, the need to clone PCR products and the sequencing of individual clones hindered the processing of large numbers of samples and the discovery of rare taxa. In 2005, 454 Life Sciences made a significant advance by producing the first true HTS platform, capable of pyrosequencing thousands to millions of individual amplified molecules in parallel (Margulies *et al.* 2005). Now, further developments in sequencing technology have further increased the depth of sequencing and opportunities for high sample throughput (Loman *et al.* 2012). In particular, Illumina sequencing-by-synthesis has enabled greater sequencing depth and higher sample throughput alongside reduced costs. More recently, single-molecule sequencing technologies, such as Pacific Biosystems and Oxford Nanopore, have allowed the generation of much longer reads from samples where DNA is only present at low concentrations: these approaches promise to be highly effective for a number of applications (e.g. genome assembly), but higher costs, reduced throughput and increased error rates mean that Illumina currently remains the platform of choice for community ecology research.

Box 1. Glossary of terms

Amplicon sequencing. Targeted sequencing of an amplified marker gene.
Community DNA. Defined here as the DNA derived from many individuals representing several species.
Cloning. The process of producing genetically identical copies of an organism, either naturally (e.g. as a result of asexual reproduction) or artificially. In the context of nucleic acid sequencing, cloning commonly refers to the insertion of DNA into a vector molecule (e.g. a plasmid) prior to selection for a gene of interest, DNA extraction and sequencing.
Degenerate primers. A mixture of similar, but not identical oligonucleotide sequences used for amplicon sequencing where the targeted gene(s) is typically similar, but not identical.
Environmental DNA (eDNA). DNA isolated directly from an environmental sample (e.g. air, faeces, sediment, soil, water).
Genomic DNA. Defined here as the DNA derived from a single individual or from a collection of individuals of the same species.
Locus. The specific location of a gene or DNA sequence on a chromosome.
Marker gene. A gene or DNA sequence targeted in amplicon sequencing to screen for a specific organism group or functional gene.
Metabarcoding. Uses gene-specific PCR primers to amplify DNA from a collection of organisms or from environmental DNA. Another term for amplicon sequencing.
Metagenomics. The random sequencing of gene fragments isolated from environmental samples, allowing sequencing of uncultivable organisms.
Metatranscriptomics. Shotgun sequencing of total RNA from environmental samples. Techniques such as poly-A amplification or rRNA depletion are often used to target messenger (mRNA) transcripts to assess gene expression patterns in complex communities.
Next generation sequencing (NGS). Recent advances in DNA sequencing that make it possible to rapidly and inexpensively sequence millions of DNA fragments in parallel. Also referred to as high-throughput sequencing (HTS).
Orthologs. Genes in different species that evolved from a common ancestor and normally retain the same function.
Polymerase chain reaction (PCR). Used to amplify a targeted piece of DNA, generating many copies of that particular DNA sequence.
Shotgun sequencing. DNA is fragmented into small segments which are individually sequenced and then reassembled into longer, continuous sequences using sequence assembly software.

Genomic, community, environmental or ancient DNA?

For the field ecologist, we can define many forms of DNA (Fig. 1). First, genomic DNA corresponds to DNA extracted from a single individual (or from a collection of individuals belonging to the same species). Secondly, community DNA consists of genomic fragments from many individuals

representing a mix of different species. Community DNA is isolated from organisms in bulk samples, but separated from their habitat (e.g. sediment, river benthos). Community DNA extracts have an important potential in ecological studies, especially for biomonitoring purposes, since the focus is on the extant community. Finally, environmental DNA (eDNA) is isolated directly from an environmental sample without first isolating any type of organism (e.g. soil, sediment, faeces, water, air) and has been the topic of many recent reviews and special issues (Taberlet *et al*. 2012b; Bohmann *et al*. 2014; Thomsen & Willerslev 2015). Environmental DNA is a complex mixture of genomic DNA from many different organisms and/or cellular material. Total eDNA from soil contains both cellular DNA and extracellular DNA (Pietramellara *et al*. 2009). Cellular DNA originates from either cells or organisms that are present within the sample and is likely to be of good quality. Extracellular DNA results from natural cell death and subsequent destruction of cell structure and is usually degraded (i.e. DNA molecules are cut into small fragments). Detecting biodiversity from eDNA was initiated and continues with a focus on prokaryotic, fungal and microeukaryotic communities, but it is now clear that we can uncover a vast amount of information about biodiversity across the three domains of life (bacteria, archaea and eukaryotes) from a broad range of source materials (Bohmann *et al*. 2014).

One of the most powerful aspects of eDNA analysis is the ability to sample biodiversity that is not easily sampled by other means or requires an unmanageable amount of time (Biggs *et al*. 2015). Contemporary eDNA analyses have already been extensively implemented for detecting invasive species in aquatic environments using species-specific markers and more recently for reliable detection of fish and/or amphibian communities (Thomsen & Willerslev 2015). In rivers, eDNA can even represent information that is integrated over large spatial areas due to the transport of DNA downstream (Deiner & Altermatt 2014). Marine sediments have also provided eDNA for analysing the pollution impact on eukaryote biodiversity in five different estuaries in Australia (Chariton *et al*. 2015). In addition to the magnitude of prokaryote studies, eDNA from soil has been used to investigate the response of soil fungi to tree dieback (Štursová *et al*. 2014), comparing plant diversity above- and below-ground (Yoccoz *et al*. 2012), shedding light on earthworm diversity (Pansu *et al*. 2015a) and cross-kingdom biodiversity assessment (Drummond *et al*. 2015). Finally, it is also possible to collect eDNA from the air as has been recently demonstrated by Kraaijeveld *et al*. (2015) using volumetric air samplers to collect pollen for allergy research.

The boundary between genomic, community and eDNA is not so precise. When isolating DNA from small organisms, the whole organism can be used. In this case, in addition to genomic DNA of the target species, the extracted DNA also contains bacterial/prey DNA from the gut and other endosymbionts. For example, when isolating DNA from a plant species, it is virtually impossible to avoid co-extracting DNA from endophytic fungi. When coring and sieving marine sediments as described in Fonseca *et al*. (2010), the resulting samples are physically enriched for meiofaunal organisms, and therefore, the extracted DNA can be considered as community DNA, but it will still contain substantial amounts of environmental DNA extracted from other organisms (e.g. undigested gut contents, or clumps of cells or tissues from larger species).

Table 1. Marker genes which are commonly used and/or recommended for marker gene assessments ('metabarcoding'). References for sequence data bases are as follows: RDP (Maidak *et al*. 1997), Greengenes (DeSantis *et al*. 2006), SILVA (Pruesse *et al*. 2007), UNITE (Abarenkov *et al*. 2010), BOLD (Ratnasingham & Hebert 2007) and Genbank (Benson *et al*. 2012).

Target	Gene/ region	Reference	Data bases
Bacteria	16S	Sogin *et al*. (2006)	RDP, Greengenes, SILVA
Archaea	16S	Sogin *et al*. (2006)	RDP, Greengenes, SILVA
Fungi	ITS	Epp *et al*. (2012), Schoch *et al*. (2012)	UNITE, GenBank, BOLD (few records)
	18S	Not recommended (Schoch *et al*. 2012)	SILVA
Protists	18S	Pawlowski *et al*. (2012)	SILVA
	ITS	Pawlowski *et al*. (2012)	GenBank
	CO1	Pawlowski *et al*. (2012)	BOLD
Meiofauna	CO1	Hebert *et al*. (2003)	BOLD
	18S	Deagle *et al*. (2014)	GenBank
Macrofauna	CO1	Hebert *et al*. (2003)	BOLD
	16S	Epp *et al*. (2012), Deagle *et al*. (2014)	GenBank
	12S	Epp *et al*. (2012), Deagle *et al*. (2014)	GenBank
	18S	Deagle *et al*. (2014)	GenBank
Plants	matK + rbcL	Hollingsworth *et al*. (2009)	GenBank, BOLD (few records)
	ITS	China Plant Barcode of Life Group (2011)	GenBank

Other highly pertinent applications of eDNA for ecologists are the study of trophic relationships using faeces as a source of eDNA (see review in Clare 2014). After the publication of a few seminal papers (Jarman, Deagle & Gales 2004; Valentini *et al*. 2009), this approach is now extensively used by ecologists for assessing the diet of herbivores (Soininen *et al*. 2015), carnivores (Deagle, Kirkwood & Jarman 2009) and omnivores (De Barba *et al*. 2014). The same forms of diet analyses have also been performed using gut contents instead of faeces (Clare 2014). In this case, even if gut contents cannot be strictly considered as eDNA, the molecular approaches are the same and yield direct insights into trophic interactions, food webs and functional ecology (Clare 2014).

Ancient eDNA represents another potent source of biodiversity information for ecologists who want to gain insights into past communities. The landmark paper of Willerslev *et al*. (2003) initiated this field of research by demonstrating that informative eDNA can be retrieved from permafrost

samples as old as 500 000 years. More recently, using hundreds of permafrost samples, Willerslev *et al.* (2014) reconstructed past plant communities in the Arctic during the last 50 000 years, based on the amplification of a short fragment of the *trnL* intron and using a large reference data base for arctic and boreal plants. Furthermore, using ancient gut contents or faeces, they were also able to determine the diet of eight individuals belonging to four herbivore species of the Quaternary megafauna (woolly mammoth, woolly rhinoceros, bison and horse). The second type of ancient eDNA exploited by ecologists is derived from lake sediments that provide a complementary tool to pollen and macrofossil analyses (Pedersen *et al.* 2013). The analysis of a 20-m-long core from a high-elevation lake in the French Alps generated the first high-resolution assessment of livestock farming history since the Neolithic perio (Giguet-Covex *et al.* 2014) and plant community trajectories over the last 6400 years (Pansu *et al.* 2015b).

Sampling approaches, preservation and DNA extraction

Environmental sequencing studies should adhere to robust ecological study design, allowing for an adequate number of sites/replicates to provide statistical power, as well as ensuring the collection of a robust set of environmental metadata (e.g. climate variables, soil pH). When designing a molecular identification protocol for detection of whole communities, there are many decisions to make. The process is linear (Fig. 1), and the steps usually consist of sample preservation, nucleic acid extraction, marker gene amplification (using PCR) or library preparation for metagenomes, sequencing the product(s) and data analysis (bioinformatics and visualization; Fig. 2). Nevertheless, the protocols used for each step can vary widely based on the question and environment (Fig. 1). The size range of the target organism typically determines how much (or little) of the physical sample is processed before DNA extraction. Microbes, viruses and other components (e.g. pollen) are easily collected from air (Kraaijeveld *et al.* 2015) and water using filtration protocols, whereby the organisms are concentrated on a series of filters with decreasing pore size that capture different size fractions of the community (Ganesh *et al.* 2014). Environmental DNA from microscopic eukaryotes is also easily captured in this way (Deiner *et al.* 2015). Cotton swabs represent another collection method used to sample microbes from animal microbiomes (e.g. skin (McKenzie *et al.* 2012)) or hard surfaces (rocks, tree bark, etc.). The effective preservation of target nucleic acids is the key starting point for any successful study. In order to preserve highly labile RNA, −80 °C temperatures and liquid nitrogen represent the gold standard, with other proprietary preservation chemicals such as RNAlater® commonly used in field sampling. DNA on the other hand is more robust and can be preserved effectively for downstream molecular biological

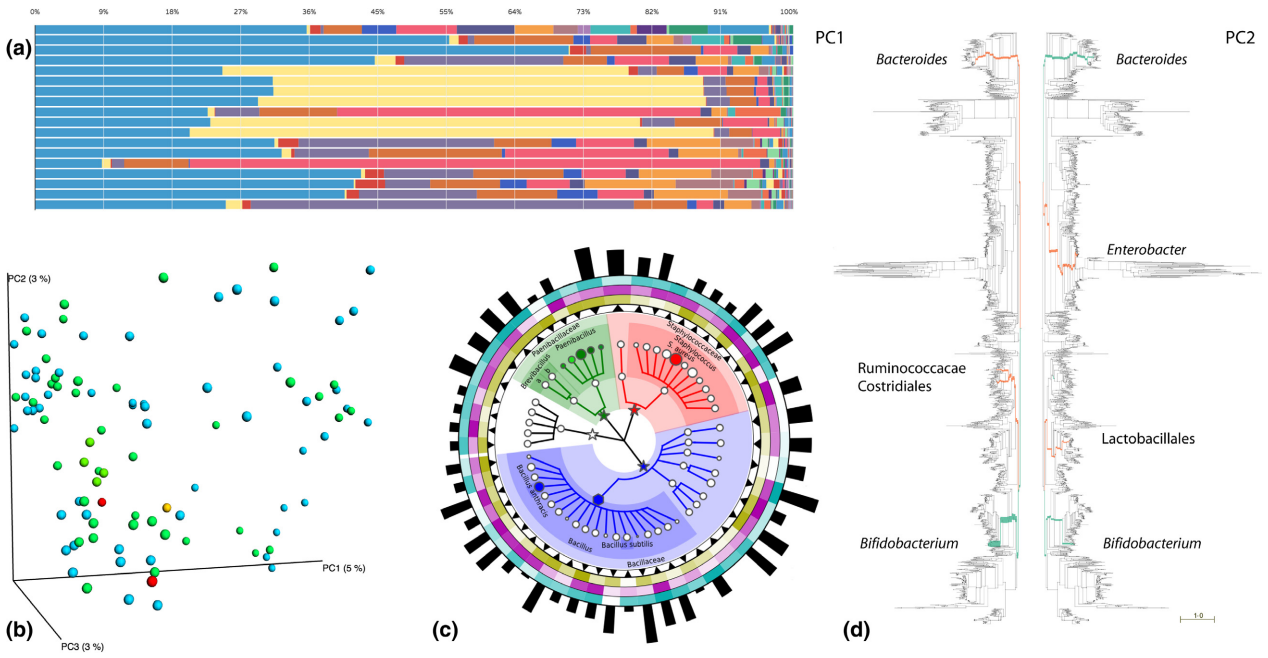

Fig. 2. Example of visualization and diversity metrics from environmental sequencing data. a) Alpha diversity displayed as taxonomy bar charts, showing relative abundance of taxa across samples using the Phinch data visualization framework (Bik & Pitch Interactive 2014). b) Beta diversity patterns illustrated via Principal Coordinate Analyses carried out in QIIME (Caporaso *et al.* 2010), where each dot represents a sample and colors distinguish different classes of sample. The closer two sample points in 3D space, the more similar their community assemblages c) GraPhalAn phylogenetic visualization of environmental data, with circular heatmaps and abundance bars used to convey quantitative taxon traits (figure reproduced from Asnicar *et al.* 2015) d) Edge PCA, a tree-based diversity metric that identifies specific lineages (green/orange branches) that contribute most to community changes observed in samples distributed across different PCA axes (Matsen & Evans 2013; figure reproduced from Darling *et al.* 2014).

manipulations by drying, −20 °C temperatures, 100% ethanol or other solutions designed to preserve both DNA and morphology such as combinations of DMSO, EDTA and saturated salt (Yoder *et al*. 2006). The use of transformed alcohols (e.g. IMS) and, in the very worst cases, formalin should be avoided since such preservation media denature nucleic acids, making them unavailable for molecular analysis.

For studies of soils and sediments, a small volume of fresh material (~0·25–2·5 g, depending on the proportion of organic matter) is typically used in DNA extraction protocols targeting microbes (e.g. bacteria/archaea, protists, fungi and viruses) (Gilbert *et al*. 2014; Pawlowski *et al*. 2014). For investigations of larger taxa such as microbial metazoa, sediments or soils are first processed via decantation/flotation protocols whereby the microbial community is separated from sediment grains (Creer *et al*. 2010). The exact method of sample processing is a critical consideration for environmental sequencing studies; any given protocol will inherently bias the view of community composition, and it is important to maintain the same protocol throughout a study in order to keep such biases consistent. Viruses and single-celled organisms are easily washed away or lysed by the decantation and sieving protocols used to isolate microbial metazoa, making it imperative to use unprocessed sediments/soils for environmental studies targeting these smaller size classes of organisms. Similarly, the low volume of fresh sediment used for DNA extractions targeting single-cell taxa does not provide sufficient material for capturing and characterizing metazoan communities. Much larger soil/sediment volumes (>100 mL) must be processed and concentrated to ensure accurate sampling for larger size classes of organisms, since microbial metazoa can exhibit spatially patchy distributions with a large number of rare species (Ramirez-Llodra *et al*. 2010). For larger organisms (e.g. macroinvertebrates), bulk communities can be homogenized or 'souped' (Yu *et al*. 2012) either with or without subsampling body parts from larger organisms that would otherwise swamp sequencing runs with excessive amounts of biomass and therefore genomic information.

Kit-based extraction protocols are an effective approach for isolating high-quality environmental DNA from microbial communities (Gilbert *et al*. 2014), although a variety of other DNA extraction methods (Griffiths *et al*. 2000; Lakay, Botha & Prior 2007) can be used depending on the scope of the study. A number of environmental studies have also used extraction approaches that enable the isolation of both DNA and RNA from a single sediment or soil sample (Griffiths *et al*. 2000; Pawlowski *et al*. 2014). In this case, RNA sequences from environmental samples from all domains of life can be revealed though reverse transcription and sequencing (McGrath *et al*. 2008). Isolation and preservation may differ from DNA methods, and this is still an area of intense research without much consensus, but rather an array of methods one can test depending on the environment sampled (De Maayer, Valverde & Cowan 2014). Co-sequencing both DNA and RNA provides, e.g. in fungi, an assessment of the 'active' community vs. potentially transient DNA from dead or inactive taxa in the environment (Baldrian *et al*. 2012).

What key methods feature in using DNA sequencing for biodiversity discovery?

MARKER GENE ASSESSMENT

Over the last decade, microbial diversity surveys have almost entirely shifted away from culture-dependent to HTS methods. Marker gene studies have become the most prevalent HTS approach, typically relying on highly degenerate PCR primers to amplify homologous taxonomy marker genes from environmental samples (Table 1). Marker gene assessments are more generally known as 'amplicon', 'metagenetic' (Creer *et al*. 2010) and 'metasystematic' (Hajibabaei 2012) sequencing among many others, but in an attempt to standardize vocabulary, nomenclature is currently converging towards the term 'metabarcoding' (Taberlet *et al*. 2012a). Metabarcoding of community DNA was first applied to marine sediments to describe meiofauna (Chariton *et al*. 2010; Creer *et al*. 2010) and subsequently to freshwater (Hajibabaei *et al*. 2011), marine (Hirai *et al*. 2015) and terrestrial (Ji *et al*. 2013) ecosystems for identifying macroinvertebrates. A consideration when choosing a marker gene locus or primer set is that not all barcodes/markers can be used to answer the same question. Different primers and gene regions vary in both taxonomic coverage and species-resolving power, leading to the introduction of taxonomic biases and associated erroneous estimates of taxon relative abundance (Bik *et al*. 2013; Klindworth *et al*. 2013). For example, standard DNA barcoding projects (e.g. the International Barcode of Life, http://ibol.org) depend on the cytochrome C oxidase subunit I as a species level diagnostic marker for animals. However, although the Barcode of Life Database (BOLD) features a standardized resource for animal identification, alternative genomic regions (e.g. nuclear 16S/18S rRNA genes, 12S mtDNA) associated with more conserved priming sites have been identified as more appropriate for 'metabarcoding' studies in certain taxa (Deagle *et al*. 2014). For example, the 18S rRNA gene exhibits extreme conservation in priming sites (Pruesse *et al*. 2007; Creer *et al*. 2010) resulting in the broad scale amplification of biodiversity across the eukaryotic tree of life, but for some taxa, this marker provides little resolving power at the species level of taxonomic resolution, even if sequenced in full using chain termination sequencing (e.g. in Fungi, ITS non-coding regions are used to resolve species) (Nilsson *et al*. 2008). Many challenges remain to species level identification of sequences, for example intraspecific variation and lack of taxonomic reference material. It should, however, be acknowledged that we do not exhaustively cover here all the challenges with this emerging technology. Continued improvements to gene maker identification of sequences from the environment are sure to follow in the years to come as the field is relatively new compared with the Linnaean system of cataloguing biodiversity.

One such solution is the multi-barcode approach (i.e. using different suites of gene markers for the same community) which has been recommended to improve taxonomic coverage and taxonomic resolution and to reduce false

negatives (Deagle *et al.* 2014; Tang *et al.* 2014). Nevertheless, the use of multiple barcodes can still be illusive as barcodes may not be equally applicable for phylogenetic vs. quantitative analyses due to primer and gene copy variation biases (See Box 1 in Supplementary information – Are the data quantitative?). An increased cost of primers, sequencing and labour are also obvious downsides of the multiple-barcode approach, in addition to the necessity of duplicity of reference data bases.

METAGENOMICS – ENVIRONMENTAL SHOTGUN SEQUENCING

Prokaryotic Communities

Although the term is often misused, true 'metagenomic' approaches utilize random sequencing of genomic fragments isolated from environmental samples to elucidate both the taxonomic and functional genomic capability of a community. In contrast to metabarcoding protocols, metagenomics can be 'PCR-free' (e.g. when using kits such as Illumina TruSeq), avoiding potential taxonomic biases stemming from use of primer sets targeting rRNA or mitochondrial loci (Logares *et al.* 2013). Shotgun sequencing can provide a complementary, independent method for assessing community diversity, additionally allowing for the capture of information from groups that are otherwise difficult to survey (Narasingarao *et al.* 2012).

Metagenomic data are typically used in two ways. The taxonomic component of shotgun sequencing can be used to identify organisms present in a sample, followed by ecologically informative alpha- and beta-diversity analyses. For example, ubiquitous loci such as rRNA genes or conserved single-copy orthologs (representing ~1% of metagenomic sequence reads) can be mined and analysed using phylogenetic workflows (Sunagawa *et al.* 2013; Darling *et al.* 2014) and tree-based metrics such as 'Edge PCA' (Fig. 2; Matsen & Evans 2013). Other approaches rely on clade-specific marker genes (if known) to classify organisms to more precise taxonomic levels (Segata *et al.* 2012). Metagenomes can also be used to characterize the functional potential of microbial communities through investigation of their full genomic repertoire. Following gene assembly, contigs are assigned putative gene functions using annotations from orthology data bases such as COG (http://www.ncbi.nlm.nih.gov/COG/) or KEGG (http://www.genome.jp/kegg/) and these categories can be compared across samples to search for potential enrichment of genes across functional classes. Alternatively, targeted gene mining approaches can be used to search for specific metabolic pathways of interest, such as nitrogen and sulphur cycling (e.g. Ganesh *et al.* 2014).

Microscopic and macroscopic eukaryotic communities

Environmental shotgun sequencing could resolve many of the issues prevalent in eukaryotic marker gene studies, particularly if it is used in conjunction with targeted genome sequencing (Fig. 3). Accordingly, the sequencing of DNA from organelles

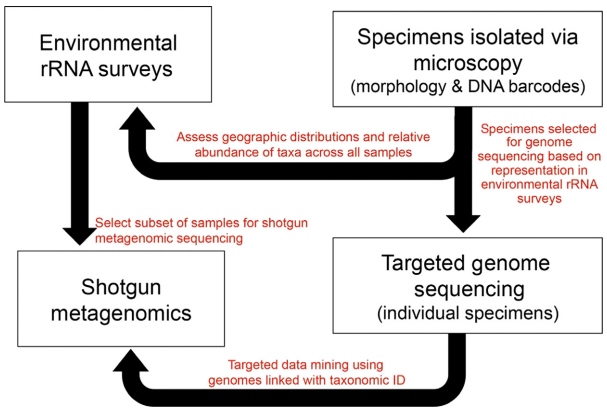

Fig. 3. Example of an integrated –Omics workflow aimed at generating genome-scale data for individual organisms alongside environmental sequencing approaches (rRNA surveys and metagenomics targeting whole communities). Assembled contigs from targeted genome sequencing can then be used to link metagenomic reads/contigs with a known taxonomic ID, via read mapping approaches.

is developing as an alternative: mitochondrial genomes for animals (Tang *et al.* 2014, 2015) and chloroplast genomes for plants (van der Merwe *et al.* 2014). For example, in chloroplast sequencing, the full genomic content of a sample is sequenced and taxonomically informative organelles are then assembled *in silico*. Focusing on shotgun-sequenced organelles, compared to particular loci, will enhance taxonomic resolution and is predicted to reduce primer/taxonomic biases, at the cost of sample throughput. Clearly, sequencing the genomes of mixed communities, compared to specific genetic loci, requires a huge increase in sequencing power and consequently a reduction in sample throughput and experimental flexibility in sampling design (Knight *et al.* 2013). A coverage/sequencing compromise relies on using DNA capture array technology to target specific organelles (Mariac *et al.* 2014). Here, arrays are designed from existing genomic organelle information which are used to hybridize and extract specific regions from genomic DNA, thereby reducing the size of the genomic target and increasing throughput. It is likely that different studies will utilize different approaches depending on budget, sample number, community composition and questions. Nevertheless, assigning taxonomy/identity to sequences derived from community DNA is implicit, and therefore, a unified stance on building specific DNA reference data bases is of utmost importance. If marker gene approaches therefore evolve into shotgun sequencing assessment of eukaryote biodiversity, we still need the ability to link genotype to phenotype. In the absence of genome sequencing all species on the planet, the utility of standardized barcode libraries (e.g. BOLD, SILVA) (Pruesse *et al.* 2007; Ratnasingham & Hebert 2007) will therefore become increasingly important and valuable to the community.

METATRANSCRIPTOMICS

Metatranscriptomics, the shotgun sequencing analysis of mRNA transcripts in environmental samples, provides near

real-time information on gene expression patterns in complex communities, that seeks to assess what gene functions of living organisms are operating at the community level. Marker gene assessment and metagenomics focus on DNA and will therefore reflect both living and dead/decaying organisms. Metatranscriptomics does not require prior knowledge of the taxonomic or functional composition of a community, and changes in mRNA transcript inventories are assumed to be indicative of activity and to provide information on the cues perceived by organisms in their environment (Gilbert & Hughes 2011; Moran *et al*. 2013). Metatranscriptomic analysis requires purification of total extracted RNA to selectively enrich for mRNA since mRNA represents a small fraction (1–5%) of the total RNA that can be extracted from environmental samples (McGrath *et al*. 2008). This can be accomplished by depleting rRNA (prokaryotes) or targeting polyA-tailed mRNA (eukaryotes). Like metagenomics, coupling the analyses of taxonomically relevant rRNA and functionally relevant mRNA provides an opportunity to link community structure and function. The approach is particularly informative for microbial communities when applied in an experimental context where both taxonomic and gene expression patterns are monitored while a particular biotic (e.g. invasive plant invasion) or abiotic (e.g. climate warming) parameter is manipulated (Moran *et al*. 2013). Environmental metatranscriptomics is not without challenges, including the inability to assign functions to a majority of mRNA sequences (existing data bases contain only genes from cultured species or the most abundant genes from a limited number of environmental samples) and the lack of a predictable relationship between mRNA abundance and protein activity (Prosser 2015). Despite these current limitations, analysis of mRNA pools in environmental samples is still a powerful omics tool for assessing microbially driven ecological processes.

SINGLE-CELL GENOMICS AND TARGETED GENOME SEQUENCING

Metagenomes, and metatranscriptomes to a lesser extent, currently represent one of the most complex types of environmental data sets we are able to generate (Howe *et al*. 2014). Nevertheless, the usefulness of shotgun data is inherently dependent on existing genome data bases for annotating contigs and inferring the functional potential of any given gene. However, for many groups of organisms – viruses, microbial metazoa and deep protist lineages in particular – there is an ongoing 'genome deficit' in which public data bases remain sparse.

It is thus not surprising that many environmental studies are increasingly using targeted genome sequencing to help to link taxonomically identified specimens with their genomic content (Fig. 3). Such approaches include traditional genome projects as well as single-cell genomics (Thrash *et al*. 2014) and computational reconstruction of abundant, small genomes from metagenomic data (prokaryotes, viruses (Sharon & Banfield 2013)). The resulting genomes can be used to assign reads from environmental shotgun data, for example to assign a taxonomic identity to 'hypothetical' proteins, or assess biogeographic patterns by mining reads from large projects such as the Global Ocean Survey (Thrash *et al*. 2014).

Bioinformatics, computational capability, infrastructure and freely available data

For clarity and succinctness, here we focus on field and laboratory approaches. Further information about the necessary bioinformatics data analysis 'tool box' can be found in the Supporting Information; this section introduces the hardware requirements, programming skills and commonly used software packages, while highlighting the freely available nature of DNA sequencing data.

The next 10 years of sequence-based meta-omic biodiversity research

We have to make ecology count and contemporary approaches should have real-life impact, including influencing policy and effectively engaging stakeholders and end-users. To this end, we have recently seen the acceptance of eDNA qPCR results to be taken as evidence of the presence of protected species in the UK (Biggs *et al*. 2015), complemented by a number of programs around the world using eDNA for the detection of alien invasive species. Metabarcoding will likely follow for high profile, costly and labour-intensive biomonitoring programs (Baird & Hajibabaei 2012), with the hope of freeing up resources to more robustly and frequently assess ecosystem health in relation to environmental stressors (Lallias *et al*. 2015). Importantly, sequencing-based approaches are not constrained to focus on particular *a priori* defined biomonitoring candidate species and may therefore yield additional insights into the interplay between environmental stressors and biodiversity of all life (Baird & Hajibabaei 2012).

Over the past 10 years, advances in sequencing technology and accompanying methodological breakthroughs have revolutionized our ability to quantify community biodiversity, but where do we go from here? From an empirical perspective, there is a clear need to link genotype to phenotype and associated ecological function (Fig. 3). There are now opportunities to map prokaryotic taxonomy marker genes to sequenced bacterial genomes of known function (Langille *et al*. 2013), complemented by metagenomics and metatranscriptomics. However, the vast task of characterizing all prokaryotic gene content will probably never be complete and the relationships between expressed mRNA transcripts and proteins/function are not always intuitive (Moran *et al*. 2013). Perhaps the biggest gains in these fields will lie in targeted assessment of specific gene pathways in relation to well-characterized systems (Toseland *et al*. 2013). From the macro-eukaryotic perspective, combinations of standardized marker gene libraries, complemented with taxonomy and metadata, do already provide a phenotype/genotype link (Ratnasingham & Hebert 2007) to functional ecology, at least as far as likely broad ecological classification, or trophic level is concerned (e.g. producer, grazer, predator, omnivore, detritivore). Therefore, these

should be supported, irrespective of the gene, or genomic approach of community biodiversity classification. As with so many studies, robust reference data bases are essential links between genes and function, including studies investigating trophic relationships/food webs (Clare 2014).

In conclusion, the standardized format and open source nature of sequencing data, accompanied by radical shifts in sequencing technology, mean that we can catalogue the spatial and temporal distribution of species from all domains of life and from all habitats. Having this global view should therefore facilitate hypothesis-driven scientific questions regarding biodiversity ecosystem–function relationships (Purdy *et al*. 2010; Hagen *et al*. 2012) in relation to external forcing, whether the drivers are anthropogenic or natural. Combined with carefully controlled experimental systems, classification of species' ecological tolerances, plasticity, distribution, rate of evolution and trophic interactions should mean that we are a step closer to making systems ecology predictions (Evans *et al*. 2013) associated with a changing environment. Without a doubt, it will certainly be challenging, but makes for exciting collaborations between the traditional fields of ecology and molecular ecologists in what is emerging to be a paradigm-shifting age of biodiversity discovery.

Acknowledgements

This manuscript resulted from in-person discussions at the 2015 joint BES/SFE meeting in Lille, France, where travel support for co-authors was provided by a US National Science Foundation Research Coordination Network award to WKT and HMB (DBI-1262480). We thank the British Ecological Society and Société Française d'Écologie for supporting our 'Ecogenomics' Symposium at this meeting. Public domain images used to construct Fig. 1 were obtained from Phylopic (http://phylopic.org/).

Data accessibility

This paper does not use any data.

References

Abarenkov, K., Henrik Nilsson, R., Larsson, K.H., Alexander, I.J., Eberhardt, U., Erland, S. *et al*. (2010) The UNITE database for molecular identification of fungi – recent updates and future perspectives. *New Phytologist*, **186**, 281–285.

Asnicar, F., Weingart, G., Tickle, T.L., Huttenhower, C. & Segata, N. (2015) Compact graphical representation of phylogenetic data and metadata with GraPhlAn. *PeerJ*, **3**, e1029.

Baird, D.J. & Hajibabaei, M. (2012) Biomonitoring 2.0: a new paradigm in ecosystem assessment made possible by next-generation DNA sequencing. *Molecular Ecology*, **21**, 2039–2044.

Baldrian, P., Kolařík, M., Štursová, M., Kopecký, J., Valášková, V., Větrovský, T. *et al*. (2012) Active and total microbial communities in forest soil are largely different and highly stratified during decomposition. *The ISME Journal*, **6**, 248–258.

Benson, D.A., Karsch-Mizrach, I.I., Clark, K., Lipman, D.J., Ostell, J. & Sayers, E.W. (2012) GenBank. *Nucleic Acids Research*, **36**, D48–D53, 201240.

Biggs, J., Ewald, N., Valentini, A., Gaboriaud, C., Dejean, T., Griffiths, R.A. *et al*. (2015) Using eDNA to develop a national citizen science-based monitoring programme for the great crested newt (*Triturus cristatus*). *Biological Conservation*, **183**, 19–28.

Bik, H.M. & Pitch Interactive (2014) Phinch: an interactive, exploratory data visualization framework for–Omic datasets. *bioRxiv*, 009944.

Bik, H.M., Porazinska, D.L., Creer, S., Caporaso, J.G., Knight, R. & Thomas, W.K. (2012) Sequencing our way towards understanding global eukaryotic biodiversity. *Trends in Ecology and Evolution*, **27**, 233–243.

Bik, H.M., Fournier, D., Sung, W., Bergeron, R.D. & Thomas, W.K. (2013) Intra-genomic variation in the ribosomal repeats of nematodes. *PLoS One*, **8**, e78230.

Blaxter, M., Mann, J., Chapman, T., Thomas, F., Whitton, C., Floyd, R. & Abebe, E. (2005) Defining operational taxonomic units using DNA barcode data. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, **360**, 1935–1943.

Bohmann, K., Evans, A., Gilbert, M.T.P., Carvalho, G.R., Creer, S., Knapp, M. *et al*. (2014) Environmental DNA for wildlife biology and biodiversity monitoring. *Trends in Ecology & Evolution*, **29**, 358–367.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K. *et al*. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*, **7**, 335–336.

Chariton, A.A., Court, L.N., Hartley, D.M., Colloff, M.J. & Hardy, C.M. (2010) Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. *Frontiers in Ecology and the Environment*, **8**, 233–238.

Chariton, A.A., Stephenson, S., Morgan, M.J., Steven, A.D.L., Colloff, M.J., Court, L.N. & Hardy, C.M. (2015) Metabarcoding of benthic eukaryote communities predicts the ecological condition of estuaries. *Environmental Pollution*, **203**, 165–174.

China Plant Barcode of Life Group, Li, D.Z., Gao, L.M., Li, H.T., Wang, H., Ge, X.J. *et al*. (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences*, **108**, 19641–19646.

Clare, E.L. (2014) Molecular detection of trophic interactions: emerging trends, distinct advantages, significant considerations and conservation applications. *Evolutionary Applications*, **7**, 1144–1157.

Creer, S., Fonseca, V.G., Porazinska, D.L., Giblin-Davis, R.M., Sung, W., Power, D.M. *et al*. (2010) Ultrasequencing of the meiofaunal biosphere: practice, pitfalls and promise. *Molecular Ecology*, **19**, 4–20.

Darling, A.E., Jospin, G., Lowe, E., Matsen, F.A. IV, Bik, H.M. & Eisen, J.A. (2014) PhyloSift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, **2**, e243.

Davies, N.Field, D. & The Genomic Observatories Network. (2012) Sequencing data: a genomic network to monitor Earth. *Nature*, **481**, 145.

De Barba, M., Miquel, C., Boyer, F., Rioux, D., Coissac, E. & Taberlet, P. (2014) DNA metabarcoding multiplexing for omnivorous diet analysis and validation of data accuracy. *Molecular Ecology Resources*, **14**, 306–323.

De Maayer, P., Valverde, A. & Cowan, D.A. (2014) The current state of metagenomic analysis. *Genome Analysis: Current Procedures and Applications*, **183**, 183–220.

Deagle, B.E., Kirkwood, R. & Jarman, S.N. (2009) Analysis of Australian fur seal diet by pyrosequencing prey DNA in faeces. *Molecular Ecology*, **18**, 2022–2038.

Deagle, B.E., Jarman, S.N., Coissac, E., Pompanon, F. & Taberlet, P. (2014) DNA metabarcoding and the cytochrome c oxidase subunit I marker: not a perfect match. *Biology Letters*, **10**, 20140562.

Deiner, K. & Altermatt, F. (2014) Transport distance of invertebrate environmental DNA in a natural river. *PLoS One*, **9**, e88786.

Deiner, K., Walser, J.-C., Mächler, E. & Altermatt, F. (2015) Choice of capture and extraction methods affect detection of freshwater biodiversity from environmental DNA. *Biological Conservation*, **183**, 53–63.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K. *et al*. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Applied and Environmental Microbiology*, **72**, 5069–5072.

Drummond, A.J., Newcomb, R.D., Buckley, T.R., Xie, D., Dopheide, A., Potter, B.C.M. *et al*. (2015) Evaluating a multigene environmental DNA approach for biodiversity assessment. *GigaScience*, **4**, 46.

Epp, L.S., Boessenkool, S., Bellemain, E.P., Haile, J., Esposito, A., Riaz, T. *et al*. (2012) New environmental metabarcodes for analysing soil DNA: potential for studying past and present ecosystems. *Molecular Ecology*, **21**, 1821–1833.

Evans, M.R., Bithell, M., Cornell, S.J., Dall, S.R.X., Diaz, S., Emmott, S. *et al*. (2013) Predictive systems ecology. *Proceedings of the Royal Society B-Biological Sciences*, **280**, 20131452.

Fonseca, V.G., Carvalho, G.R., Sung, W., Johnson, H.F., Power, D.M., Neill, S.P. *et al*. (2010) Second-generation environmental sequencing unmasks marine metazoan biodiversity. *Nature Communications*, **1**, 98.

Fox, G.E., Stackebrandt, E., Hespell, R.B., Gibson, J., Maniloff, J., Dyer, T.A. *et al*. (1980) The phylogeny of prokaryotes. *Science*, **209**, 457–463.

Ganesh, S., Parris, D.J., DeLong, E.F. & Stewart, F.J. (2014) Metagenomic analysis of size-fractionated picoplankton in a marine oxygen minimum zone. *The ISME Journal*, **8**, 187–211.

Giguet-Covex, C., Pansu, J., Arnaud, F., Rey, P.J., Griggo, C., Gielly, L. *et al.* (2014) Long livestock farming history and human landscape shaping revealed by lake sediment DNA. *Nature Communications*, **5**, 3221.

Gilbert, J.A. & Hughes, M. (2011) Gene expression profiling: metatranscriptomics. *High-Throughput Next Generation Sequencing: Methods and Application* (eds Y.M. Kwon & C. Ricke), pp. 95–205. Humana Press, New York.

Gilbert, J.A., Jansson, J.K. & Knight, R. (2014) The Earth Microbiome project: successes and aspirations. *BMC Biology*, **12**, 69.

Giovannoni, S.J., Britschgi, T.B., Moyer, C.L. & Field, K.G. (1990) Genetic diversity in Sargasso Sea bacterioplankton. *Nature*, **345**, 60–63.

Griffiths, R.I., Whiteley, A.S., O'Donnell, A.G. & Bailey, M.J. (2000) Rapid method for coextraction of DNA and RNA from natural environments for analysis of ribosomal DNA- and rRNA-based microbial community composition. *Applied and Environmental Microbiology*, **66**, 5488–5491.

Group, C.P.W., Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M. & Ratnasingham, S. *et al.* (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, **106**, 12794–12797.

Hagen, M., Kissling, W.D., Rasmussen, C., Carstensen, D.W., Dupont, Y.L., Kaiser-Bunbury, C.N. *et al.* (2012) Biodiversity, species interactions and ecological networks in a fragmented world. *Advances in Ecological Research*, **46**, 89–120.

Hajibabaei, M. (2012) The golden age of DNA metasystematics. *Trends in Genetics*, **28**, 535–537.

Hajibabaei, M., Shokralla, S., Zhou, X., Singer, G.A.C. & Baird, D.J. (2011) Environmental barcoding: a next-generation sequencing approach for biomonitoring applications using river benthos. *PLoS One*, **6**, e17497.

Hebert, P.D.N., Ratnasingham, S. & de Waard, J.R. (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. *Proceedings of the Royal Society of London B: Biological Sciences*, **270**, S96–S99.

Hirai, J., Kuriyama, M., Ichikawa, T., Hidaka, K. & Tsuda, A. (2015) A metagenetic approach for revealing community structure of marine planktonic copepods. *Molecular Ecology Resources*, **15**, 68–80.

Hollingsworth, P.M., Forrest, L.L., Spouge, J.L., Hajibabaei, M., Ratnasingham, S., van der Bank, M. *et al.* (2009) A DNA barcode for land plants. *Proceedings of the National Academy of Sciences*, **106**, 12794–12797.

Howe, A.C., Jansson, J.K., Malfatti, S.A., Tringe, S.G., Tiedje, J.M. & Brown, C.T. (2014) Tackling soil diversity with the assembly of large, complex metagenomes. *Proceedings of the National Academy of Sciences USA*, **111**, 4904–4909.

Jarman, S.N., Deagle, B.E. & Gales, N.J. (2004) Group-specific polymerase chain reaction for DNA-based analysis of species diversity and identity in dietary samples. *Molecular Ecology*, **13**, 1313–1322.

Ji, Y.Q., Ashton, L., Pedley, S.M., Edwards, D.P., Tang, Y., Nakamura, A. *et al.* (2013) Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. *Ecology Letters*, **16**, 1245–1257.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M. & Gloeckner, F.O. (2013) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Research*, **41**, e1.

Knight, R., Jansson, J.K., Field, D., Fierer, N., Desai, N., Fuhrman, J.A. *et al.* (2013) Unlocking the potential of metagenomics through replicated experimental design. *Nature Biotechnology*, **30**, 513–520.

Kraaijeveld, K., de Weger, L., Garcia, M.V., Buermans, H., Frank, J., Hiemstra, P.S. & den Dunnen, J. (2015) Efficient and sensitive identification and quantification of airborne pollen using next-generation DNA sequencing. *Molecular Ecology Resources*, **15**, 8–16.

Lakay, F.M., Botha, A. & Prior, B.A. (2007) Comparative analysis of environmental DNA extraction and purification methods from different humic acid-rich soils. *Journal of Applied Microbiology*, **102**, 265–273.

Lallias, D., Geert Hiddink, J., Fonseca, V.G., Gaspar, J.M., Sung, W., Neill, S.P. *et al.* (2015) Environmental meta-barcoding reveals heterogeneous drivers of microbial eukaryotic diversity in contrasting estuarine ecosystems. *The ISME Journal*, **9**, 1208–1221.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A. *et al.* (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nature Biotechnology*, **31**, 814–821.

Lawson Handley, L.J. (2015) How will the 'molecular revolution' contribute to biological recording? *Biological Journal of the Linnean Society*, **115**, 750–766.

Logares, R., Sunagawa, S., Salazar, G., Cornejo-Castillo, F.M., Ferrera, I., Sarmento, H. *et al.* (2013) Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environmental Microbiology*, **16**, 2659–2671.

Loman, N.J., Misra, R.V., Dallman, T.J., Constantinidou, C., Gharbia, S.E., Wain, J. & Pallen, M.J. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, **30**, 434–439.

Maidak, B.L., Olsen, G.J., Larsen, N., Overbeek, R., McCaughey, M.J. & Woese, C.R. (1997) The ribosomal database project (RDP). *Nucleic Acids Research*, **24**, 82–85.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A. *et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376–380.

Mariac, C., Scarcelli, N., Pouzadou, J., Barnaud, A., Billot, C., Faye, A. *et al.* (2014) Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. *Molecular Ecology Resources*, **14**, 1103–1113.

Matsen, F.A. IV & Evans, S.N. (2013) Edge principal components and squash clustering: using the special structure of phylogenetic placement data for sample comparison. *PLoS One*, **8**, e56859.

McGrath, K.C., Thomas-Hall, S.R., Cheng, C.T., Leo, L., Alexa, A., Schmidt, S. & Schenk, P.M. (2008) Isolation and analysis of mRNA from environmental microbial communities. *Journal of Microbiological Methods*, **75**, 172–176.

McKenzie, V.J., Bowers, R.M., Fierer, N., Knight, R. & Lauber, C.L. (2012) Co-habiting amphibian species harbor unique skin bacterial communities in wild populations. *The ISME Journal*, **6**, 588–596.

van der Merwe, M., McPherson, H., Siow, J. & Rossetto, M. (2014) Next-Gen phylogeography of rainforest trees: exploring landscape-level cpDNA variation from whole-genome sequencing. *Molecular Ecology Resources*, **14**, 199–208.

Moran, M.A., Satinsky, B., Gifford, S.M., Luo, H., Rivers, A., Chan, L.-K. *et al.* (2013) Sizing up metatranscriptomics. *The ISME Journal*, **7**, 237–243.

Narasingarao, P., Podell, S., Ugalde, J.A., Brochier-Armanet, C., Emerson, J.B., Brocks, J.J. *et al.* (2012) *De novo* metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. *The ISME Journal*, **6**, 81–93.

Nilsson, R.H., Kristiansson, E., Ryberg, M., Hallenberg, N. & Larsson, K.H. (2008) Intraspecific ITS variability in the kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evolutionary Bioinformatics*, **4**, 193–201.

Pansu, J., De Danieli, S., Puissant, J., Gonzalez, J.M., Gielly, L., Cordonnier, T. *et al.* (2015a) Landscape-scale distribution patterns of earthworms inferred from soil DNA. *Soil Biology and Biochemistry*, **83**, 100–105.

Pansu, J., Giguet-Covex, C., Ficetola, G.F., Gielly, L., Boyer, F., Zinger, L. *et al.* (2015b) Reconstructing long-term human impacts on plant communities: an ecological approach based on lake sediment DNA. *Molecular Ecology*, **24**, 1485–1498.

Pawlowski, J., Audic, S., Adl, S., Bass, D., Belbahri, L., Berney, C. *et al.* (2012) CBOL protist working group: barcoding eukaryotic richness beyond the animal, plant, and fungal kingdoms. *PLoS Biology*, **10**, e1001419.

Pawlowski, J., Esling, P., Lejzerowicz, F., Cedhagen, T. & Wilding, T.A. (2014) Environmental monitoring through protist next-generation sequencing metabarcoding: assessing the impact of fish farming on benthic foraminifera communities. *Molecular Ecology Resources*, **14**, 1129–1140.

Pedersen, M.W., Ginolhac, A., Orlando, L., Olsen, J., Andersen, K., Holm, J. *et al.* (2013) A comparative study of ancient environmental DNA to pollen and macrofossils from lake sediments reveals taxonomic overlap and additional plant taxa. *Quaternary Science Reviews*, **75**, 161–168.

Pietramellara, G., Ascher, J., Borgogni, F., Ceccherini, M.T., Guerri, G. & Nannipieri, P. (2009) Extracellular DNA in soil and sediment: fate and ecological relevance. *Biology and Fertility of Soils*, **45**, 219–235.

Prosser, J.I. (2015) Dispersing misconceptions and identifying opportunities for the use of 'omics' in soil microbial ecology. *Nature Reviews Microbiology*, **13**, 439–446.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B., Ludwig, W., Peplies, J. & Glockner, F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Research*, **35**, 7188–7196.

Purdy, K.J., Hurd, P.J., Moya-Laraño, J., Trimmer, M., Oakley, B.B. & Woodward, G. (2010) Systems biology for ecology: from molecules to ecosystems. *Advances in Ecological Research*, **43**, 87–149.

Ramirez, K.S., Leff, J.W., Barberan, A., Bates, S.T., Betley, J., Crowther, T.W. *et al.* (2014) Biogeographic patterns in below-ground diversity in New York City's Central Park are similar to those observed globally. *Proceedings of the Royal Society B*, **281**, 1795.

Ramirez-Llodra, E., Brandt, A., Danovaro, R., De Mol, B., Escobar, E., German, C.R. *et al.* (2010) Deep, diverse and definitely different: unique attributes of the world's largest ecosystem. *Biogeosciences*, **7**, 2851–2899.

Ratnasingham, S. & Hebert, P.D.N. (2007) BOLD: the barcode of life data system (www.barcodinglife.org). *Molecular Ecology Notes*, **7**, 355–364.
Saiki, R.K., Gelfland, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T. *et al.* (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**, 487–491.
Schoch, C.L., Seifert, K.A., Huhndorf, S., Robert, V., Spouge, J.L., Levesque, C.A. *et al.* (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences*, **109**, 6241–6246.
Segata, N., Waldron, L., Ballarini, A., Narasimhan, V., Jousson, O. & Huttenhower, C. (2012) Metagenomic microbial community profiling using unique clade-specific marker genes. *Nature Methods*, **9**, 811–814.
Segata, N., Boernigen, D., Tickle, T.L., Morgan, X.C., Garrett, W.S. & Huttenhower, C. (2013) Computational meta'omics for microbial community studies. *Molecular Systems Biology*, **9**, 666.
Sharon, I. & Banfield, J.F. (2013) Genomes from metagenomics. *Science*, **342**, 1057–1058.
Soininen, E.M., Gauthier, G., Bilodeau, F., Berteaux, D., Gielly, L., Taberlet, P. *et al.* (2015) Highly overlapping winter diet in two sympatric lemming species revealed by DNA metabarcoding. *PLoS One*, **10**, e0115335.
Sogin, M.L., Morrison, H.G., Huber, J.A., Welch, D.W., Huse, S.M., Neal, P.R., Arrieta, J.M. & Herndl, G.J. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences*, **103**, 12115–12120.
Štursová, M., Šnajdr, J., Cajthaml, T., Bárta, J., Šantrůčková, H. & Baldrian, P. (2014) When the forest dies: the response of forest soil fungi to a bark beetle-induced tree dieback. *The ISME Journal*, **8**, 1920–1931.
Sunagawa, S., Mende, D.R., Zeller, G., Izquierdo-Carrasco, F., Berger, S.A., Kultima, J.R. *et al.* (2013) Metagenomic species profiling using universal phylogenetic marker genes. *Nature Methods*, **10**, 1196.
Taberlet, P., Coissac, E., Hajibabaei, M. & Rieseberg, L.H. (2012a) Environmental DNA. *Molecular Ecology*, **21**, 1789–1793.
Taberlet, P., Coissac, E., Pompanon, F., Brochmann, C. & Willerslev, E. (2012b) Towards next-generation biodiversity assessment using DNA metabarcoding. *Molecular Ecology*, **21**, 2045–2050.
Tang, M., Tan, M., Meng, G., Yang, S., Su, X., Liu, S. *et al.* (2014) Multiplex sequencing of pooled mitochondrial genomes – a crucial step toward biodiversity analysis using mito-metagenomics. *Nucleic Acids Research*, **42**, e166.
Tang, M., Hardman, C.J., Ji, Y., Meng, G., Liu, S., Tan, M. *et al.* (2015) High-throughput monitoring of wild bee diversity and abundance via mitogenomics. *Methods in Ecology and Evolution*, **6**, 1034–1043.
Thomsen, P.F. & Willerslev, E. (2015) Environmental DNA – an emerging tool in conservation for monitoring past and present biodiversity. *Biological Conservation*, **183**, 4–18.
Thrash, J.C., Temperton, B., Swan, B.K., Landry, Z.C., Woyke, T., DeLong, E.F. *et al.* (2014) Single-cell enabled comparative genomics of a deep ocean SAR11 bathytype. *The ISME Journal*, **8**, 1440–1451.
Toseland, A., Daines, S.J., Clark, J.R., Kirkham, A., Strauss, J., Uhlig, C. *et al.* (2013) The impact of temperature on marine phytoplankton resource allocation and metabolism. *Nature Climate Change*, **3**, 979–984.
Valentini, A., Miquel, C., Nawaz, M.A., Bellemain, E., Coissac, E., Pompanon, F. *et al.* (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach. *Molecular Ecology Resources*, **9**, 51–60.
Vellend, M. (2010) Conceptual synthesis in community ecology. *Quarterly Review of Biology*, **85**, 183–206.
Willerslev, E., Hansen, A.J., Binladen, J., Brand, T.B., Gilbert, M.T.P., Shapiro, B. *et al.* (2003) Diverse plant and animal genetic records from Holocene and Pleistocene sediments. *Science*, **300**, 791–795.
Willerslev, E., Davison, J., Moora, M., Zobel, M., Coissac, E., Edwards, M.E. *et al.* (2014) Fifty-five thousand years of arctic vegetation and megafaunal diet. *Nature*, **506**, 47–51.
Yoccoz, N.G., Bråthen, K.A., Gielly, L., Haile, J., Edwards, M.E., Goslar, T. *et al.* (2012) DNA from soil mirrors plant taxonomic and growth form diversity. *Molecular Ecology*, **21**, 3647–3655.
Yoder, M., Ley, I.T.D., King, I.W., Mundo-Ocampo, M., Mann, J., Blaxter, M. *et al.* (2006) DESS: a versatile solution for preserving morphology and extractable DNA of nematodes. *Nematology*, **8**, 367–376.
Yu, D.W., Ji, Y., Emerson, B.C., Wang, X., Ye, C., Yang, C. & Ding, Z. (2012) Biodiversity soup: metabarcoding of arthropods for rapid biodiversity assessment and biomonitoring. *Methods in Ecology and Evolution*, **3**, 613–623.

Received 11 December 2015; accepted 30 March 2016
Handling Editor: Robert Freckleton

Supporting Information

Additional Supporting Information may be found online in the supporting information tab for this article:

Appendix S1. Bioinformatics and computational capability.
Box 1. Are the data quantitative?